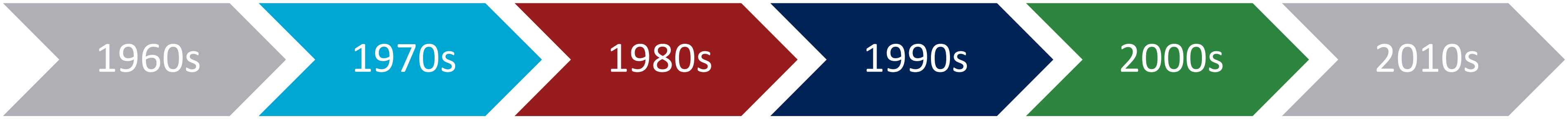
NIH Virtual Workshop: Ultra Large Chemistry Databases

Chemical Space is Infinite: How can one scale to infinity while still being usable/useful?

Evan Bolton, Ph.D. - Program Head of Chemistry

NIH U.S. National Library of Medicine
National Center for Biotechnology Information

An evolution of chemical structure databases



1960s

- Computers aid publication process
- Computerization of chemistry

1970s

- Chemical structure publisher databases (Beilstein, CAS)
- Rise of databases

1980s

- Chemical structure corporate databases
- Dial-up accessibility (STN)
- Computer savvy chemists

1990s

- Desktop chemical structure databases (ISIS BASE)
- Internet accessible chemical structure databases (SciFinder)
- Large private collections of information

2000s

- Internet accessible, open chemical structure databases you can also download and use locally
- Rise of open chemistry

2010s

- Large open chemical databases of aggregated content
- RESTful programmatic access

Special thanks to discussions with Dr. John Rumble and Dr. Evan Hepler-Smith

This is what it means to be useful for many use-cases

Imagine you wanted to make a modern scientific resource, what do you need to focus on?

Use of persistent research identifiers

Data use cases explicitly considered

Standards-based approaches

Explicit data licensing (e.g., CC-BY 4.0, CC0)

2020s Cloud-first, Mobile-first, Machine-first, FAIR-first, Open-first

Receive cloud-based data

Make data accessible within cloud

UI/UX from a device screen size agnostic perspective

Use of controlled vocabulary and machine interpretable statements

Sufficient meta data to reproduce the science

FAIR means “**F**ully **AI** **R**eady” .. also means “**F**indable”, “**A**ccessible”, “**I**nteroperable”, “**R**eusable”

Are we ready for ultra-large chemical databases?

- InChIKey, will it still work?
- Users, will they know how to use?
- Is interactivity a thing of the past?
- Are the possibilities so great that it all just seems random?

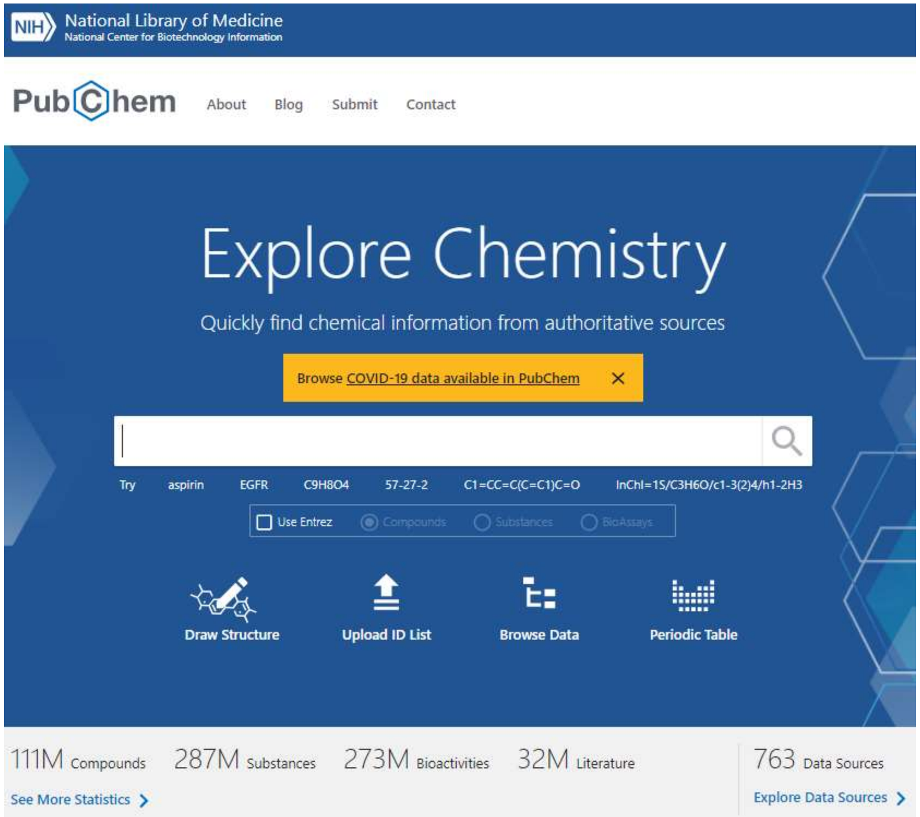
Image credit:
https://ilekh.com/wp-content/uploads/2014/08/time-for-change.jpg

What if PubChem was 1 billion structures? 10 billion structures? Would it be more or less useful?

PubChem is a data repository

- World's largest collection of freely accessible chemical information.
- Helps researchers make sense of the biological roles and health effects of chemicals on human health and the environment.

Chemical substances and bioactivities .. with select annotation



Explore Chemistry

Quickly find chemical information from authoritative sources

Browse COVID-19 data available in PubChem ×

Try aspirin EGFR C9H8O4 57-27-2 C1=CC=C(C=C1)C=O InChI=1S/C3H6O/c1-3(2)4/h1-2H3

□ Use Entrez ● Compounds ○ Substances ○ BioAssays

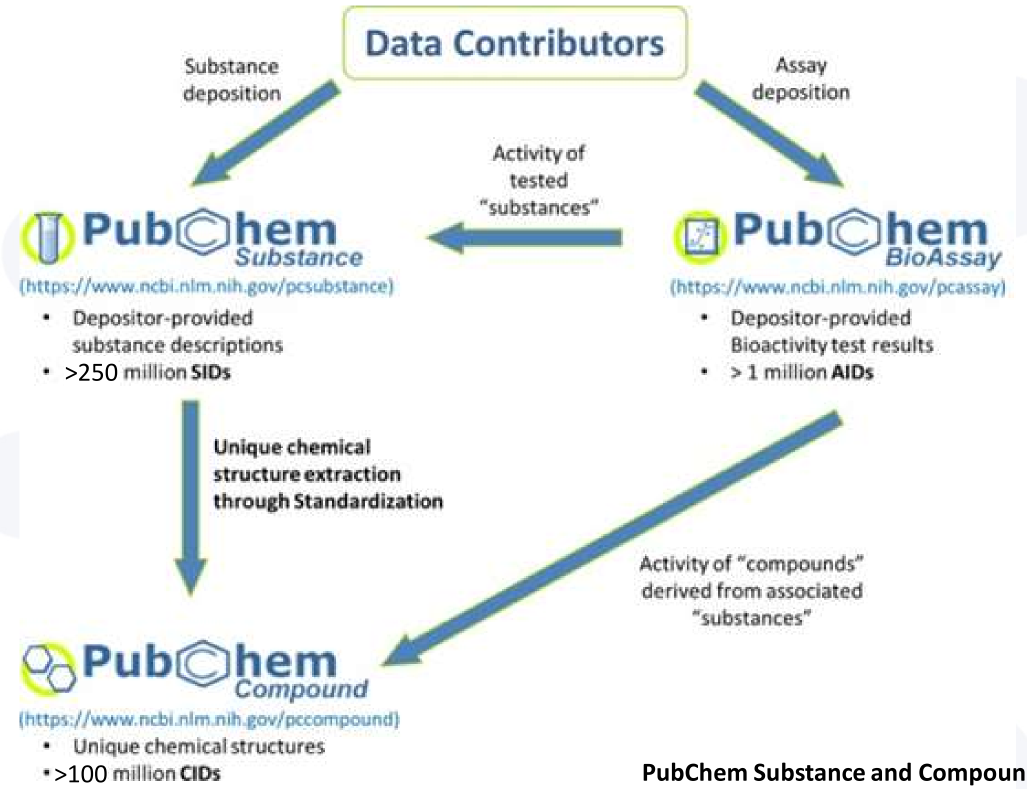
111M Compounds 287M Substances 273M Bioactivities 32M Literature 763 Data Sources

https://pubchem.ncbi.nlm.nih.gov/

U.S. National Library of Medicine
National Center for Biotechnology Information

Lots of links between records exist between +250M bioactivities, +250M substances, +100M compounds, +1M bioassays

Two primary archival databases



Data Contributors

Substance deposition

Assay deposition

Activity of tested "substances"

PubChem Substance

(https://www.ncbi.nlm.nih.gov/pcsubstance)

- Depositor-provided substance descriptions
- >250 million **SIDs**

PubChem BioAssay

(https://www.ncbi.nlm.nih.gov/pcassay)

- Depositor-provided Bioactivity test results
- > 1 million **AIDs**

Unique chemical structure extraction through Standardization

Activity of "compounds" derived from associated "substances"

PubChem Compound

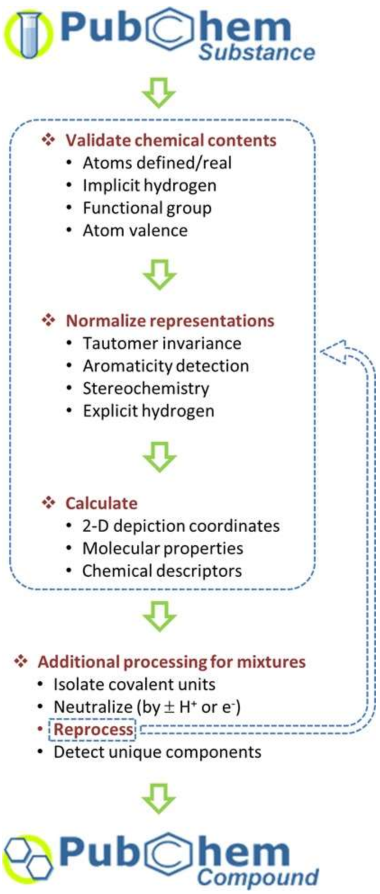
(https://www.ncbi.nlm.nih.gov/pccompound)

- Unique chemical structures
- >100 million **CIDs**

U.S. National Library of Medicine
National Center for Biotechnology Information

PubChem Substance and Compound databases
Nucleic Acids Res. 2016 Jan 4;44(D1):D1202-13.
DOI: 10.1093/nar/gkv951
PMID: 26400175
PMCID: PMC4702940

Compound is derived from Substance



PubChem Substance

- Validate chemical contents
  - Atoms defined/real
  - Implicit hydrogen
  - Functional group
  - Atom valence
- Normalize representations
  - Tautomer invariance
  - Aromaticity detection
  - Stereochemistry
  - Explicit hydrogen
- Calculate
  - 2-D depiction coordinates
  - Molecular properties
  - Chemical descriptors
- Additional processing for mixtures
  - Isolate covalent units
  - Neutralize (by ± H^+ or e^-)
  - Reprocess
  - Detect unique components

PubChem Compound

Many to many links abound in PubChem within a given collection

Many to many relationships

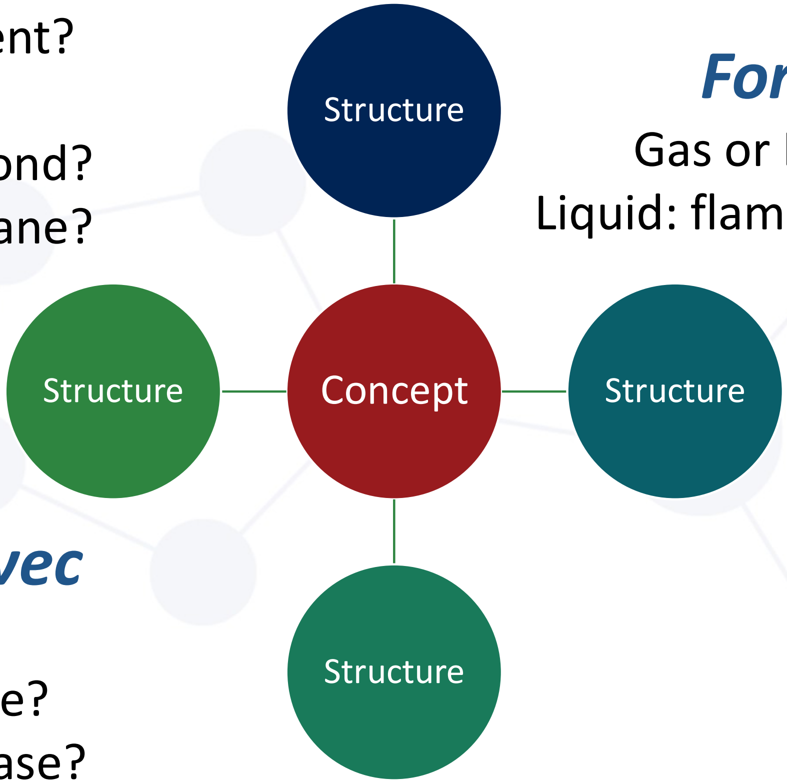
Carbon

Element?

Coal?

Diamond?

Methane?



Gleevec

Salt?

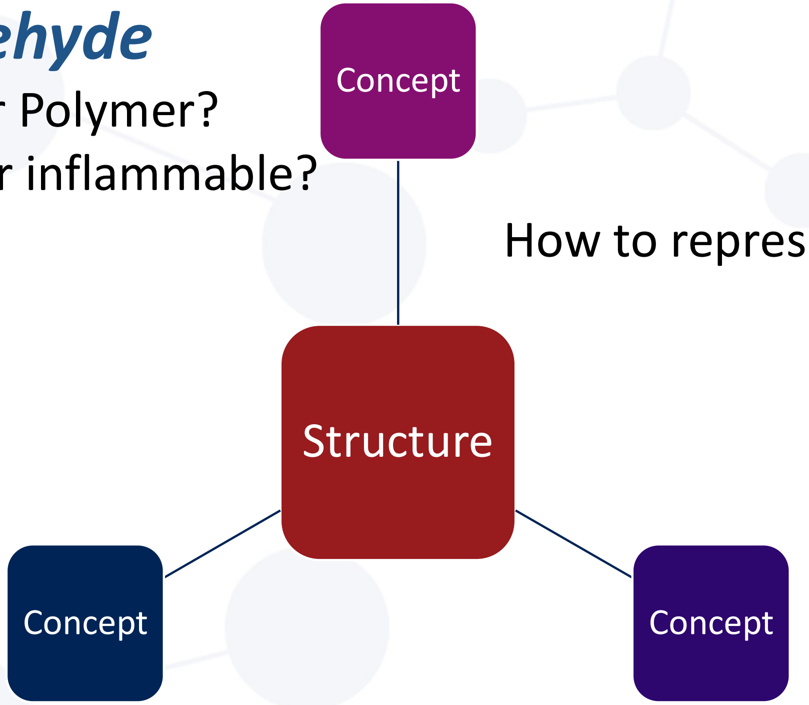
Hydrate?

Free base?

Formaldehyde

Gas or Liquid or Polymer?

Liquid: flammable or inflammable?

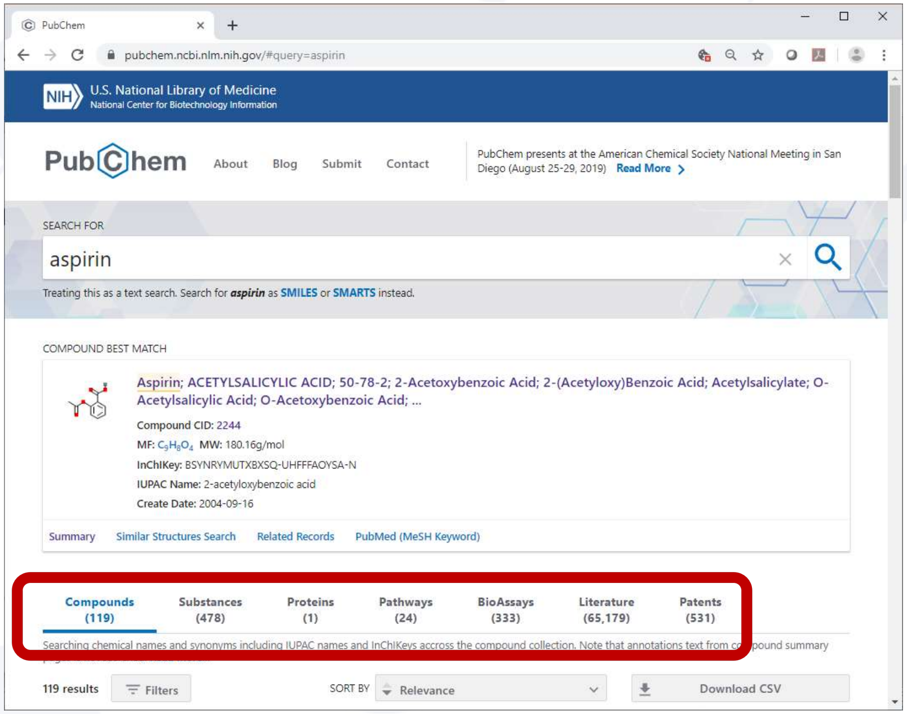


How to represent?

Many links between other external collections and annotation

Integrating many resources

- Over 700 contributing resources
  - Unified search engine
    - Flexible handling of query types
    - Includes chemical input extensions and sketcher
  - Many collection types
    - Compounds / Substances
    - Proteins / Genes / *Pathways*
    - BioAssays
    - Literature / Patents


PubChem
pubchem.ncbi.nlm.nih.gov/#query=aspirin
NIH U.S. National Library of Medicine
National Center for Biotechnology Information
PubChem
About
Blog
Submit
Contact

PubChem presents at the American Chemical Society National Meeting in San Diego (August 25-29, 2019) Read More ›

SEARCH FOR

aspirin

Treating this as a text search. Search for ***aspirin*** as **SMILES** or **SMARTS** instead.

COMPOUND BEST MATCH

Aspirin; ACETYLSALICYLIC ACID; 50-78-2; 2-Acetoxybenzoic Acid; 2-(Acetyloxy)Benzoic Acid; Acetylsalicylate; O-Acetylsalicylic Acid; O-Acetoxybenzoic Acid; ...

Compound CID: 2244

MF: $C_9H_8O_4$ MW: 180.16g/mol

InChIKey: BSYNRYMUTXBXSQ-UHFFFAOYSA-N

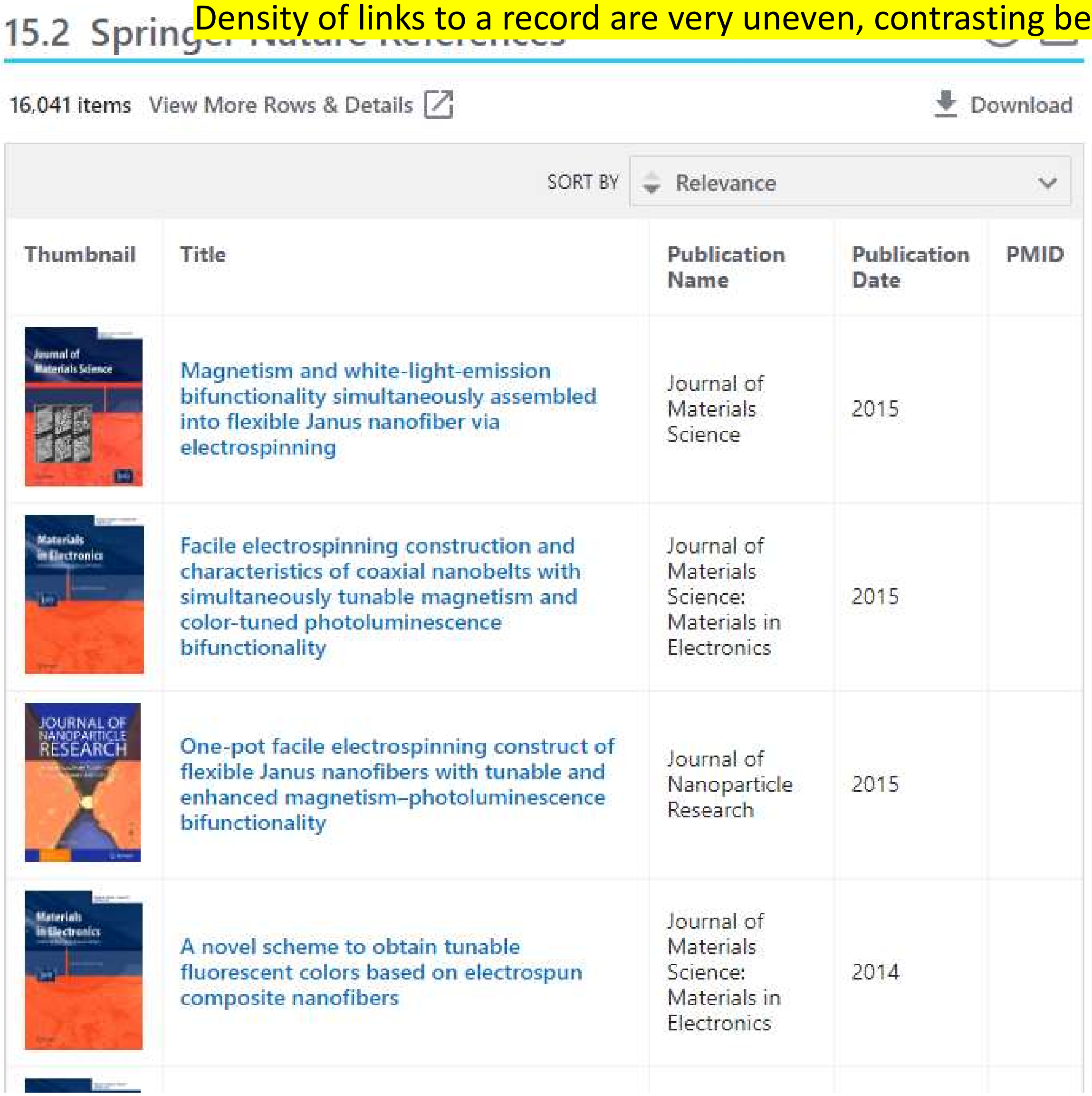
IUPAC Name: 2-acetyloxybenzoic acid

Create Date: 2004-09-16

Searching chemical names and synonyms including IUPAC names and InChIKeys accross the compound collection. Note that annotations text from compound summary

119 results

Density of links to a record are very uneven, contrasting between very sparse and very dense



15.2 Springer Nature References

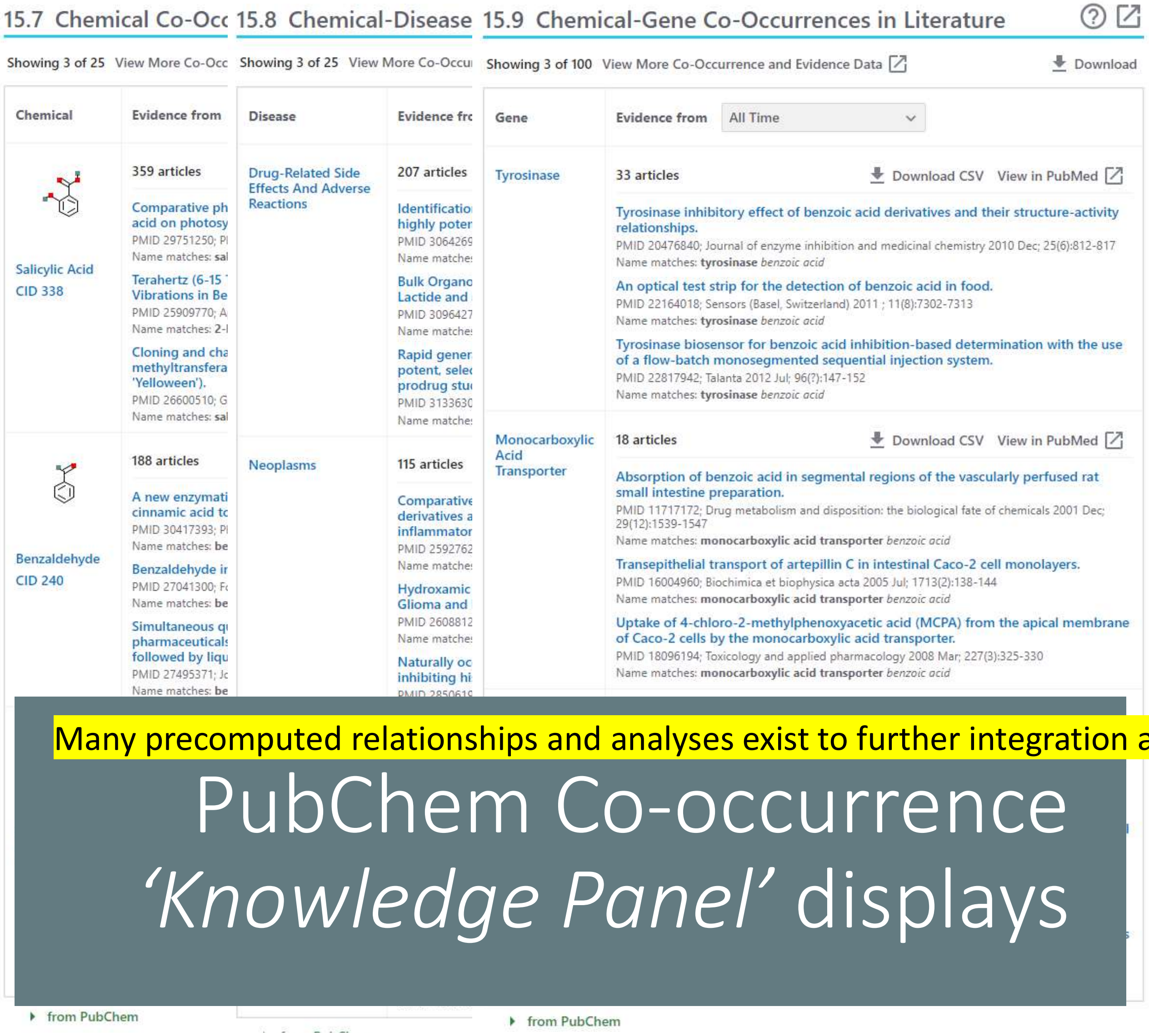
16,041 items View More Rows & Details

Thumbnail	Title	Publication Name	Publication Date	PMID
	Magnetism and white-light-emission bifunctionality simultaneously assembled into flexible Janus nanofiber via electrospinning	Journal of Materials Science	2015	
	Facile electrospinning construction and characteristics of coaxial nanobelts with simultaneously tunable magnetism and color-tuned photoluminescence bifunctionality	Journal of Materials Science: Materials in Electronics	2015	
	One-pot facile electrospinning construct of flexible Janus nanofibers with tunable and enhanced magnetism–photoluminescence bifunctionality	Journal of Nanoparticle Research	2015	
	A novel scheme to obtain tunable fluorescent colors based on electrospun composite nanofibers	Journal of Materials Science: Materials in Electronics	2014	

Integrating publisher provided metadata

- Major publishers provide chemical-DOI associations
  - Thieme
  - Springer Nature
- Publishers provide document level metadata
  - CrossRef, PubMed, SciGraph, (Agricola)

PubChem Co-occurrence *'Knowledge Panel'* displays



15.7 Chemical Co-Occ

Showing 3 of 25 View More Co-Occ

Chemical	Evidence from
Salicylic Acid CID 338	359 articles Comparative ph acid on photosy PMID 29751250; P Name matches: **sal** Terahertz (6-15 Vibrations in Be PMID 25909770; A Name matches: **2-** Cloning and cha methyltransfera 'Yelloween'). PMID 26600510; G Name matches: **sal**
Benzaldehyde CID 240	188 articles A new enzymati cinnamic acid tc PMID 30417393; P Name matches: **be** Benzaldehyde ir PMID 27041300; F Name matches: **be** Simultaneous q pharmaceuticals followed by liqu PMID 27495371; J Name matches: **be**

from PubChem

15.8 Chemical-Disease

Showing 3 of 25 View More Co-Occu

Disease	Evidence fro
Drug-Related Side Effects And Adverse Reactions	207 articles Identificatio highly poter PMID 3064269 Name matche Bulk Organo Lactide and PMID 3096427 Name matche Rapid gener potent, selec prodrug stu PMID 3133630 Name matche
Neoplasms	115 articles Comparative derivatives a inflammator PMID 2592762 Name matche Hydroxamic Glioma and PMID 2608812 Name matche Naturally oc inhibiting hi PMID 2850619

15.9 Chemical-Gene Co-Occurrences in Literature

Showing 3 of 100 View More Co-Occurrence and Evidence Data

Download

Gene	Evidence from All Time
Tyrosinase	33 articles Download CSV View in PubMed Tyrosinase inhibitory effect of benzoic acid derivatives and their structure-activity relationships. PMID 20476840; Journal of enzyme inhibition and medicinal chemistry 2010 Dec; 25(6):812-817 Name matches: **tyrosinase** *benzoic acid* An optical test strip for the detection of benzoic acid in food. PMID 22164018; Sensors (Basel, Switzerland) 2011 ; 11(8):7302-7313 Name matches: **tyrosinase** *benzoic acid* Tyrosinase biosensor for benzoic acid inhibition-based determination with the use of a flow-batch monosegmented sequential injection system. PMID 22817942; Talanta 2012 Jul; 96(?):147-152 Name matches: **tyrosinase** *benzoic acid*
Monocarboxylic Acid Transporter	18 articles Download CSV View in PubMed Absorption of benzoic acid in segmental regions of the vascularly perfused rat small intestine preparation. PMID 11717172; Drug metabolism and disposition: the biological fate of chemicals 2001 Dec; 29(12):1539-1547 Name matches: **monocarboxylic acid transporter** *benzoic acid* Transepithelial transport of artepillin C in intestinal Caco-2 cell monolayers. PMID 16004960; Biochimica et biophysica acta 2005 Jul; 1713(2):138-144 Name matches: **monocarboxylic acid transporter** *benzoic acid* Uptake of 4-chloro-2-methylphenoxyacetic acid (MCPA) from the apical membrane of Caco-2 cells by the monocarboxylic acid transporter. PMID 18096194; Toxicology and applied pharmacology 2008 Mar; 227(3):325-330 Name matches: **monocarboxylic acid transporter** *benzoic acid*

from PubChem

- Using PubMed corpus
- Mine text title/abstract
- Find all chemical, disease, gene/protein mentions
- Compute histogram and provide top-N
- Evidence clearly stated
  - name, PMIDs, ..
  - Downloadable
- Handles all nine combinations of Chemical, Gene/Protein, Disease

Many precomputed relationships and analyses exist to further integration and interpretation by users and machines

The chemical information $\overset{\text{understanding}}{\vee}$ divide

Human understanding

Depictions
Schemes
Table
Text

Computer understanding

Explicit
Complete
Annotated
Interpreted

Image credit:
https://hcldr.files.wordpress.com/2016/12/1-rr-0116-industry-divided.jpg

The richness of data is further enhanced by leveraging machine-based semantic relationships

PubChem RDF-based Linked Data

- Describes relationships between PubChem data
- Machine-readable information triples
- Organized like layers of an onion (just take what you need)
- Uses ontologies and vocabulary description
- Only a core set of PubChem content
  - Expanding coverage

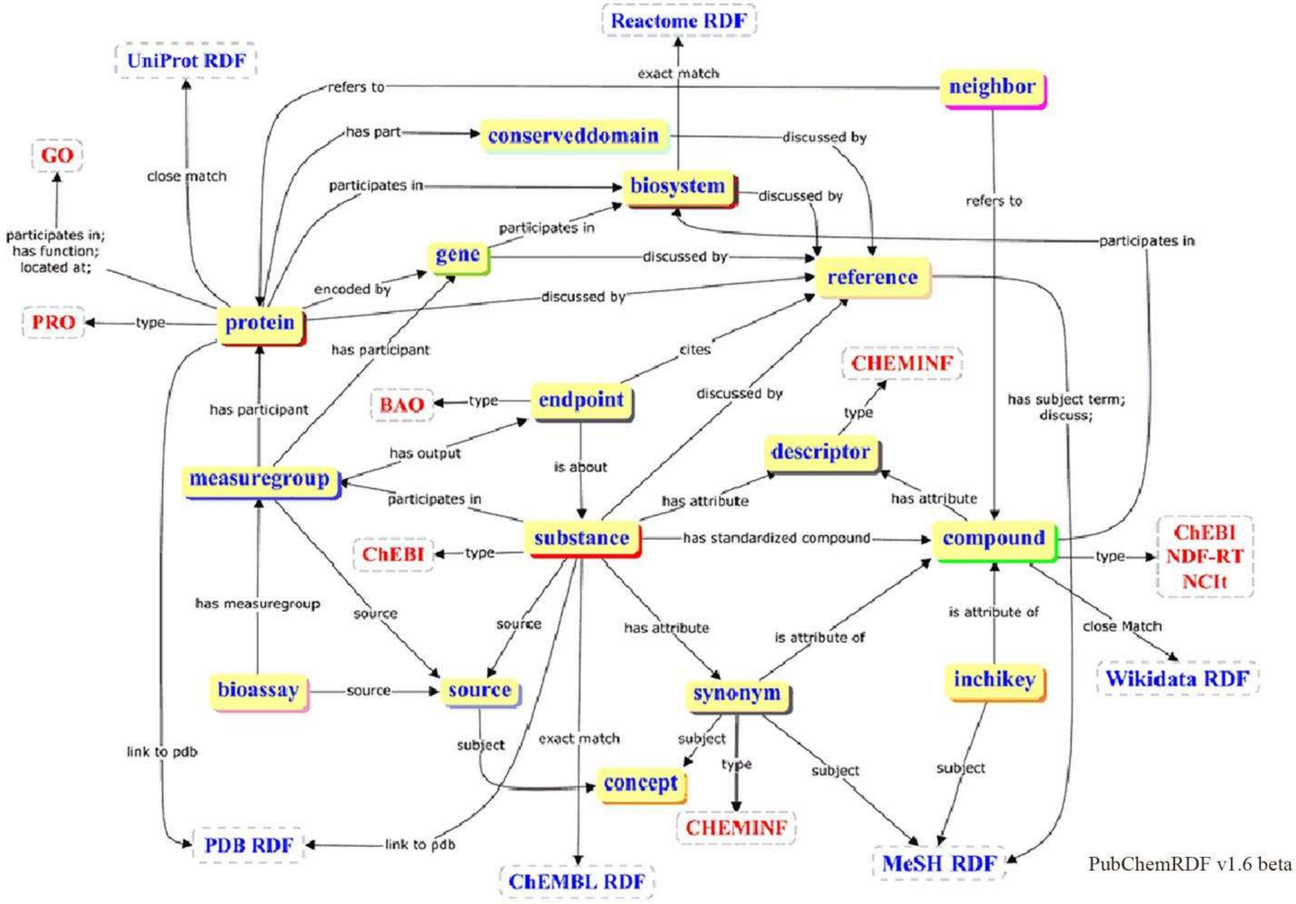


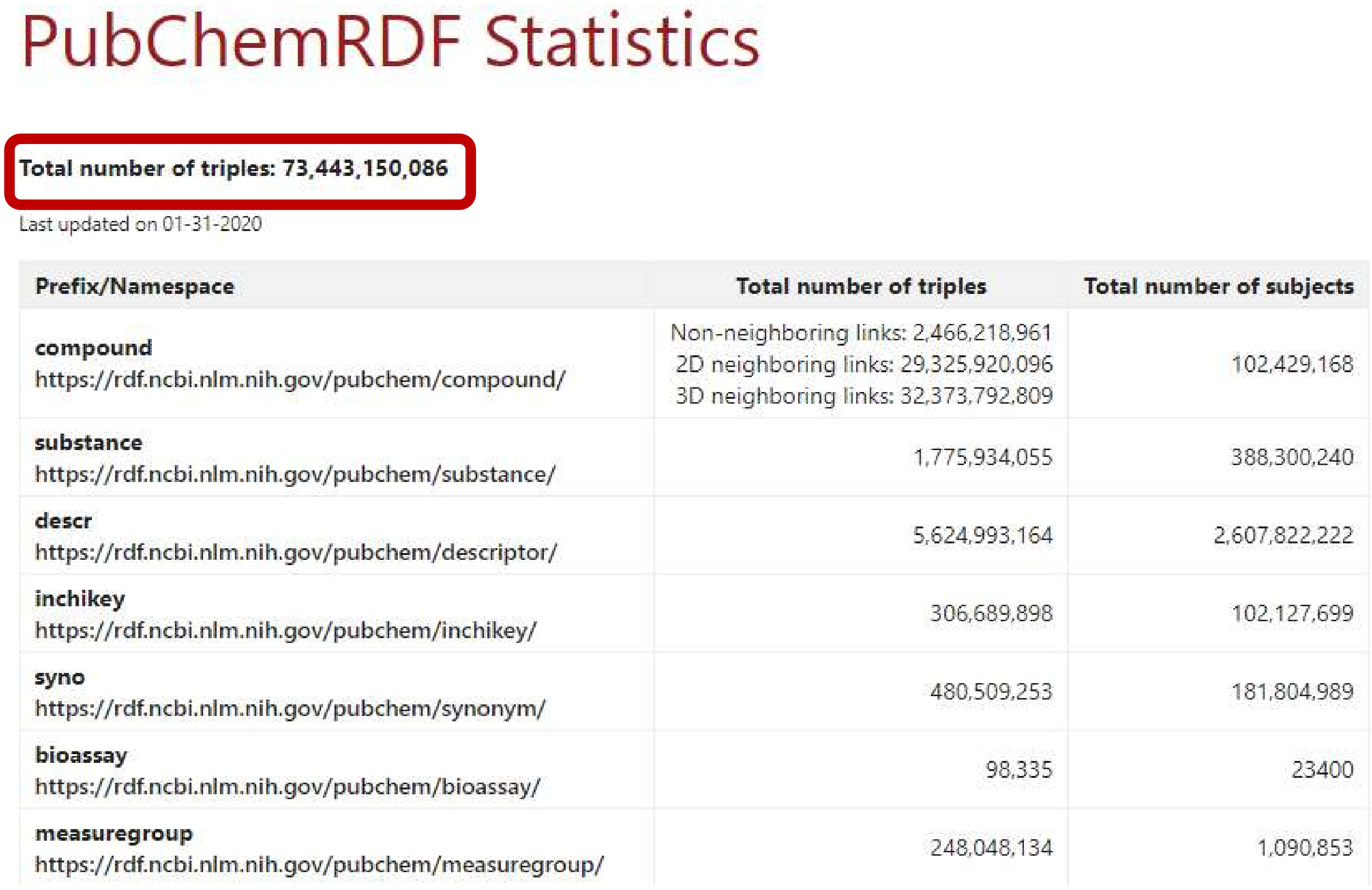
Figure 1. Color-coded diagram showing a high-level overview of the PubChemRDF semantic relationships.

U.S. National Library of Medicine
National Center for Biotechnology Information

PubChem limits the degree of links within the RDF representation to keep the count under 100 billion .. could be 10x larger

PubChem RDF-based Linked Data

- Describes relationships between PubChem data
- Machine-readable information triples
- Organized like layers of an onion (just take what you need)
- Uses ontologies and vocabulary description
- Only a core set of PubChem content
- Working to add more



PubChemRDF Statistics

Total number of triples: 73,443,150,086

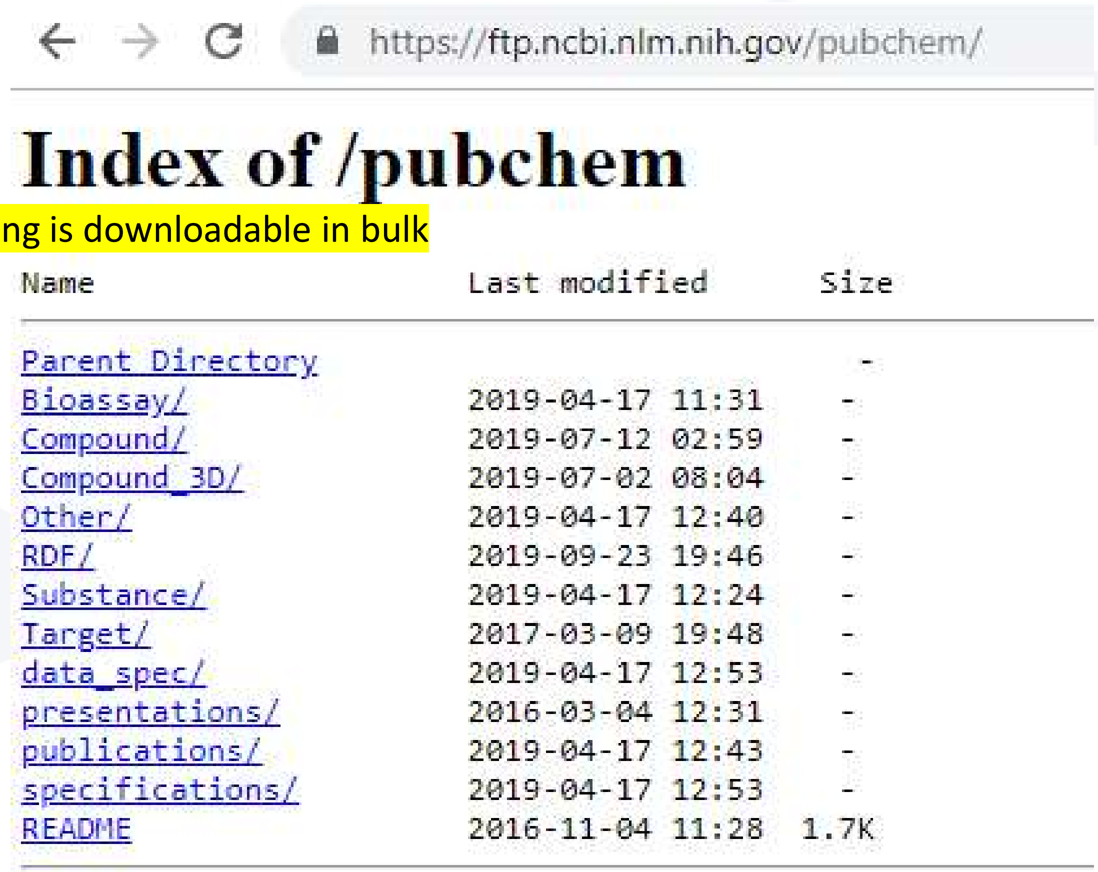
Last updated on 01-31-2020

Prefix/Namespace	Total number of triples	Total number of subjects
compound https://rdf.ncbi.nlm.nih.gov/pubchem/compound/	Non-neighboring links: 2,466,218,961 2D neighboring links: 29,325,920,096 3D neighboring links: 32,373,792,809	102,429,168
substance https://rdf.ncbi.nlm.nih.gov/pubchem/substance/	1,775,934,055	388,300,240
descr https://rdf.ncbi.nlm.nih.gov/pubchem/descriptor/	5,624,993,164	2,607,822,222
inchikey https://rdf.ncbi.nlm.nih.gov/pubchem/inchikey/	306,689,898	102,127,699
syno https://rdf.ncbi.nlm.nih.gov/pubchem/synonym/	480,509,253	181,804,989
bioassay https://rdf.ncbi.nlm.nih.gov/pubchem/bioassay/	98,335	23400
measuregroup https://rdf.ncbi.nlm.nih.gov/pubchem/measuregroup/	248,048,134	1,090,853

… and more …

PubChem FTP Site

- Different file formats
- 7 TBs of data
- Compound, Substance, BioAssay, Bioactivities
- Lots of extras
  - Special data sets
  - Link files to patents, PubMed, …
  - Target
  - RDF
  - Specifications



https://ftp.ncbi.nlm.nih.gov/pubchem/

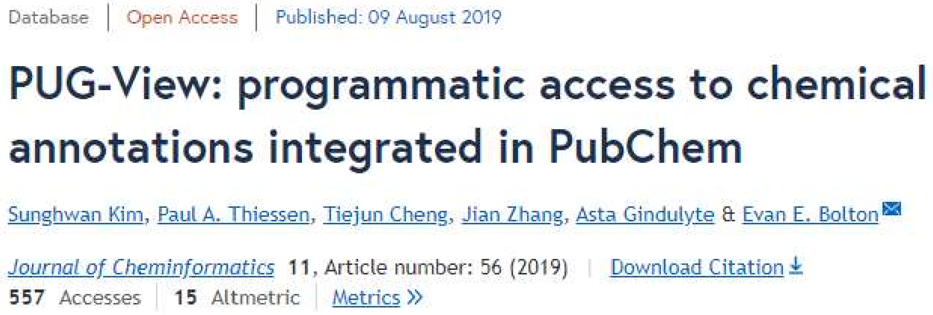
Index of /pubchem

Everything is downloadable in bulk

Name	Last modified	Size
Parent Directory		-
Bioassay/	2019-04-17 11:31	-
Compound/	2019-07-12 02:59	-
Compound_3D/	2019-07-02 08:04	-
Other/	2019-04-17 12:40	-
RDF/	2019-09-23 19:46	-
Substance/	2019-04-17 12:24	-
Target/	2017-03-09 19:48	-
data_spec/	2019-04-17 12:53	-
presentations/	2016-03-04 12:31	-
publications/	2019-04-17 12:43	-
specifications/	2019-04-17 12:53	-
README	2016-11-04 11:28	1.7K

ftp://ftp.ncbi.nlm.nih.gov/pubchem/
https://ftp.ncbi.nlm.nih.gov/pubchem/

Annotation-based data access


Database | Open Access | Published: 09 August 2019

PUG-View: programmatic access to chemical annotations integrated in PubChem

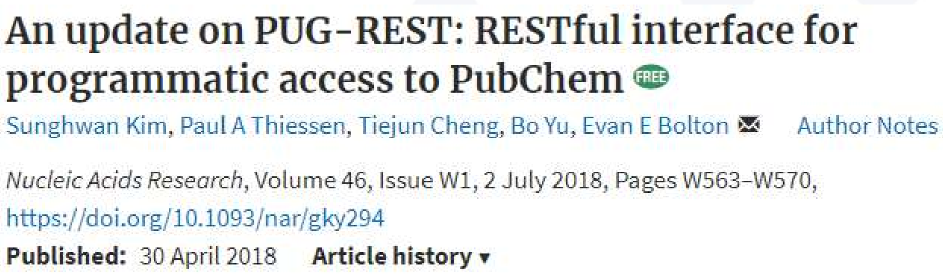
Sunghwan Kim, Paul A. Thiessen, Tiejun Cheng, Jian Zhang, Asta Gindulyte & Evan E. Bolton

Journal of Cheminformatics 11, Article number: 56 (2019) | Download Citation

557 Accesses | 15 Altmetric | Metrics »

PMID: 31399858
PMCID: PMC6324075
DOI: 10.1186/s13321-019-0375-2

Archive-based data access


An update on PUG-REST: RESTful interface for programmatic access to PubChem

Sunghwan Kim, Paul A Thiessen, Tiejun Cheng, Bo Yu, Evan E Bolton   Author Notes

Nucleic Acids Research, Volume 46, Issue W1, 2 July 2018, Pages W563–W570, https://doi.org/10.1093/nar/gky294

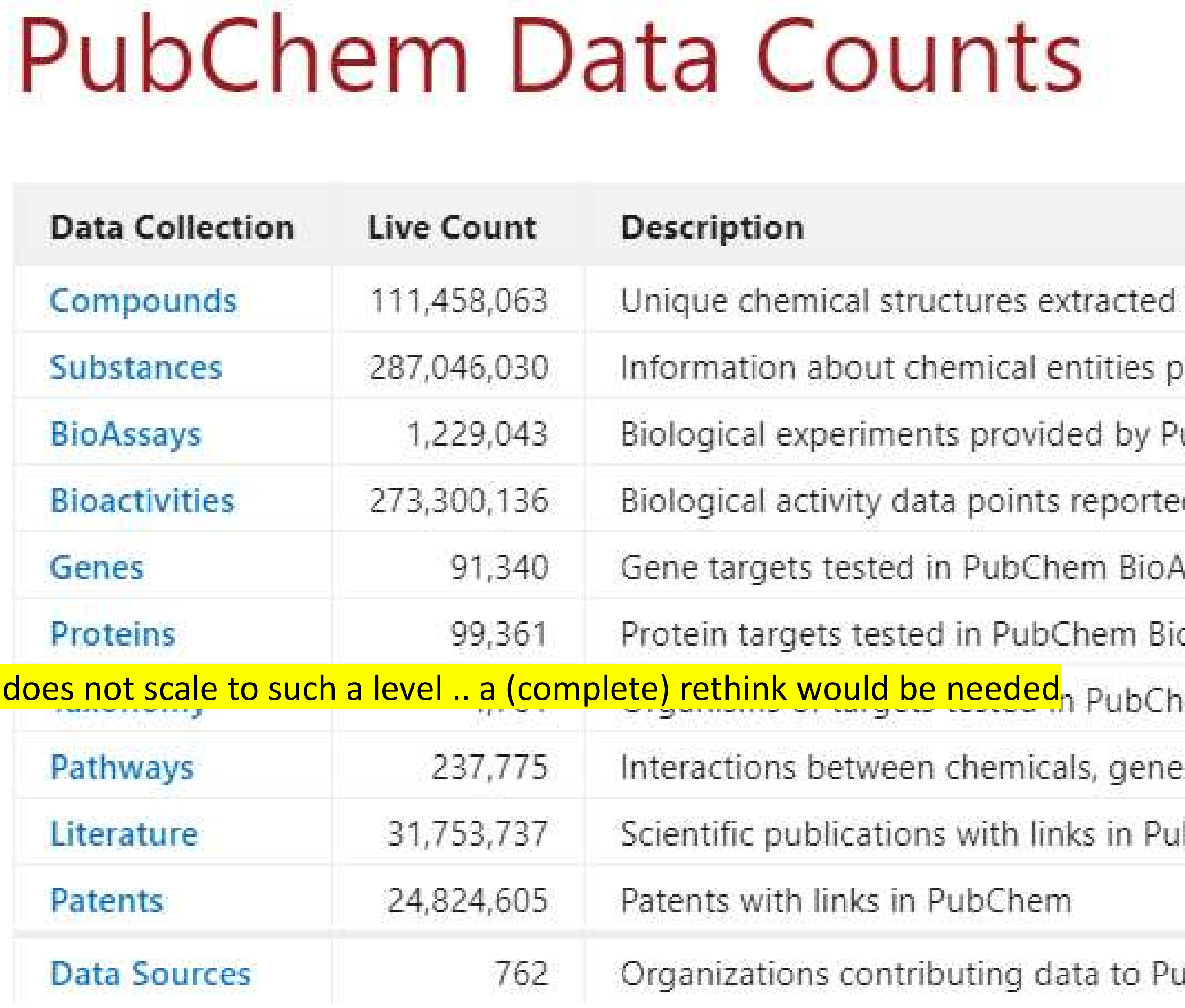
Published: 30 April 2018   **Article history** ▾

PMID: 29718389
PMCID: PMC6030920
DOI: 10.1093/nar/gky294

Content is programmatically accessible

Everything is analyzable or downloadable in pieces interactively and programmatically .. capacity is needed at scale

What would happen if all these counts increased 10x?



PubChem Data Counts

Data Collection	Live Count	Description
Compounds	111,458,063	Unique chemical structures extracted
Substances	287,046,030	Information about chemical entities p
BioAssays	1,229,043	Biological experiments provided by Pu
Bioactivities	273,300,136	Biological activity data points reporte
Genes	91,340	Gene targets tested in PubChem BioA
Proteins	99,361	Protein targets tested in PubChem Bio
Taxonomy	[illegible]	Organisms of targets tested in PubCh
Pathways	237,775	Interactions between chemicals, gene
Literature	31,753,737	Scientific publications with links in Pu
Patents	24,824,605	Patents with links in PubChem
Data Sources	762	Organizations contributing data to Pu

Short answer .. PubChem would implode .. it does not scale to such a level .. a (complete) rethink would be needed

Contemplating chemical infinity

What holds us back from chemical infinity?

One or more of the following:

1. Hardware
2. Software
3. Money
4. Time
5. Use cases

Image credit:
https://www.masterdc.com/images/photos/dedicated-and-managed-server-differences.jpg

Can we find better ways to scale?

Very basic operations of a chemical structure database

- Query by structure
  - **Identity** – scales as N
  - **Similarity** – scales as N
  - **Substructure** – scales as N-ish
    (depends on the algorithm and tradeoffs)
- **Sort** results – scales as N * ln N
- **Filter** results – scales as N
- **Retrieve** results – scales as N

(assumes a full scan needed, many optimizations can be applied, completely ignores analysis beyond filtering)

Considering only

Database Size	N * ln N
1M	10M
10M	100M
100M	1B
1B	10B
10B	100B
100B	1T
1T	10T
10T	100T

M=million, B=billion, T=trillion

Can we implement more memory efficient approaches?

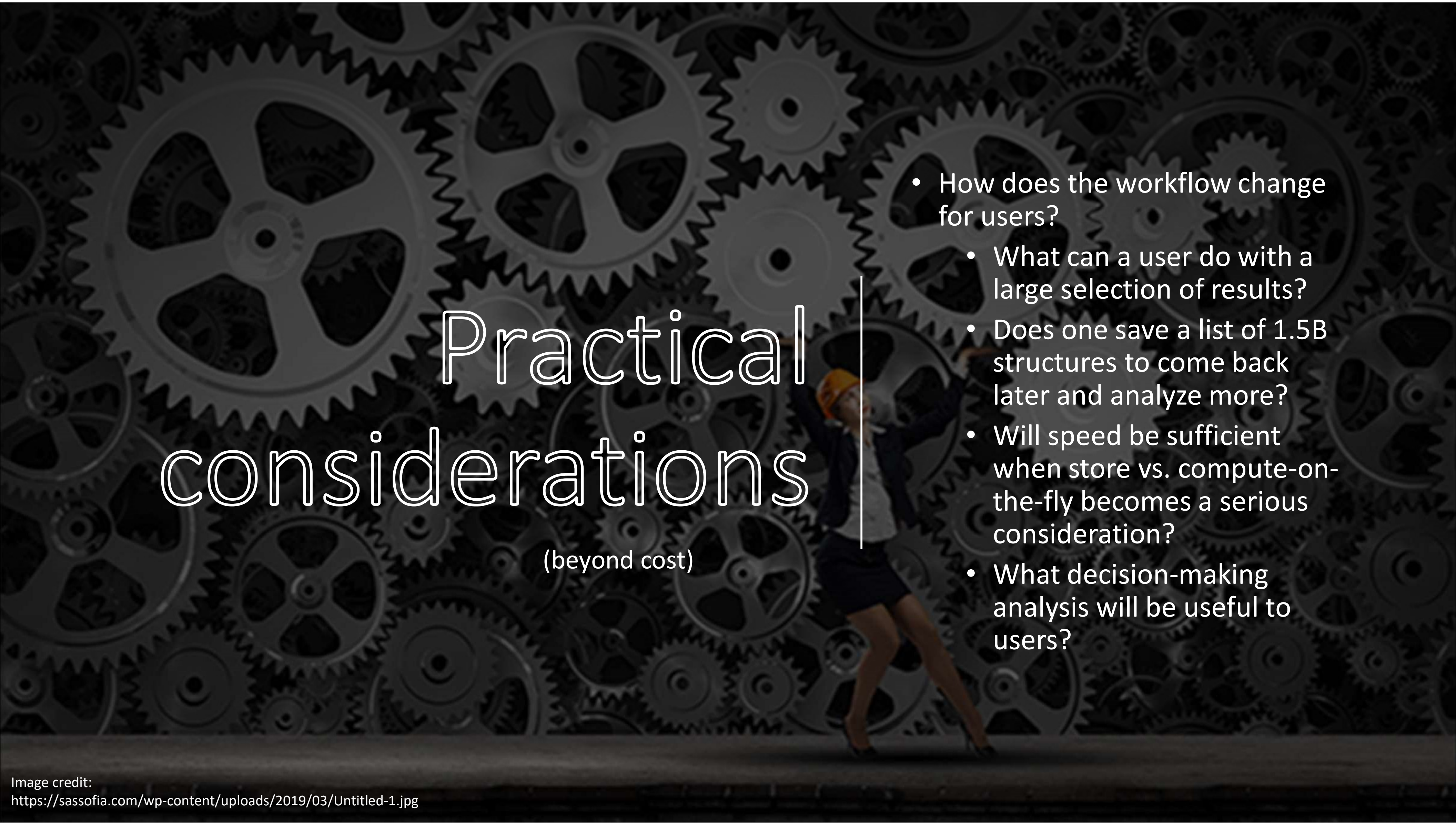
M=million, B=billion, T=trillion
MB=megabyte, GB=gigabyte, PB=petabyte, EB=exabyte

Consider the economics of a feature-less chemical structure database in the cloud

- Assuming
  - 100 bytes per structure
  - vCPU w/ 2GB memory @ $0.025/hour
  - Storage 1GB @ $0.08/month
  - Minimal I/O used (else $$)
  - All in memory (for speed!)
  - Capacity for only one query at a time

Database Size	Storage	vCPU needed	Yearly cost
1M	100MB	1	$ 219
10M	1GB	1	$ 220
100M	10GB	5	$ 1,105
1B	100GB	50	$ 11,046
10B	1PB	500	$ 110,460
100B	10PB	5,000	$ 1,104,600
1T	100PB	50,000	$ 11,046,000
10T	1EB	500,000	$ 110,460,000

(many tradeoffs can be applied to change the economics, ignores many aspects needed to run at scale, a full featured chemical search system likely 10-100 times more expensive before optimization, analysis costs extra)

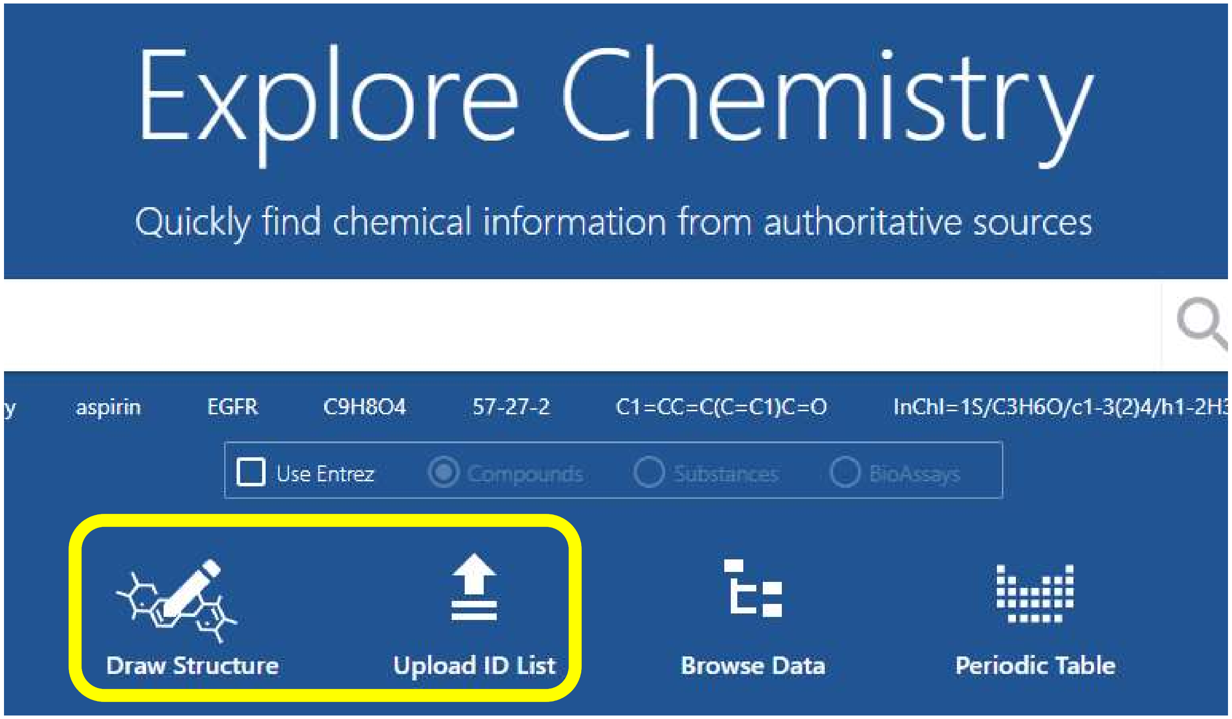


Practical considerations

(beyond cost)

- How does the workflow change for users?
  - What can a user do with a large selection of results?
  - Does one save a list of 1.5B structures to come back later and analyze more?
  - Will speed be sufficient when store vs. compute-on-the-fly becomes a serious consideration?
  - What decision-making analysis will be useful to users?

Image credit:
https://sassofia.com/wp-content/uploads/2019/03/Untitled-1.jpg

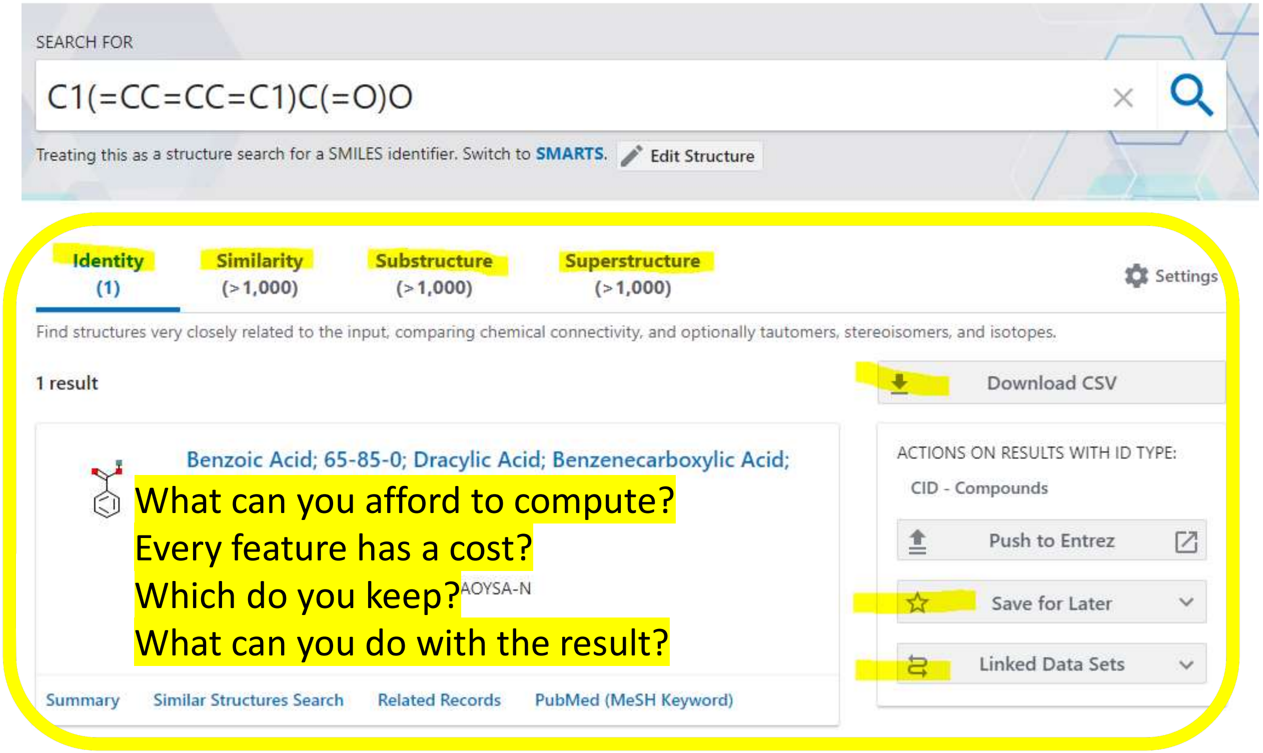


How do you import data?
How much can you import?
How do you use large inputs on database side?

PubChem Search

https://pubchem.ncbi.nlm.nih.gov/

- Single box, many query types
  - Chemical name, CAS#
  - Gene symbol/name
  - Molecular Formula
  - SMILES, InChI
  - SMARTS (substructure)
  - …
- Draw a structure
  - Or upload a file
- Chemical Search by
  - Identity, similarity, substructure, superstructure, mol. formula
- Upload an ID list
- Many collections
  - Compounds, Substances, BioAssays, Genes, Proteins, Pathways, Literature, Patents



SEARCH FOR

C1(=CC=CC=C1)C(=O)O

Treating this as a structure search for a SMILES identifier. Switch to **SMARTS**. Edit Structure

Find structures very closely related to the input, comparing chemical connectivity, and optionally tautomers, stereoisomers, and isotopes.

1 result

Benzoic Acid; 65-85-0; Dracylic Acid; Benzenecarboxylic Acid;

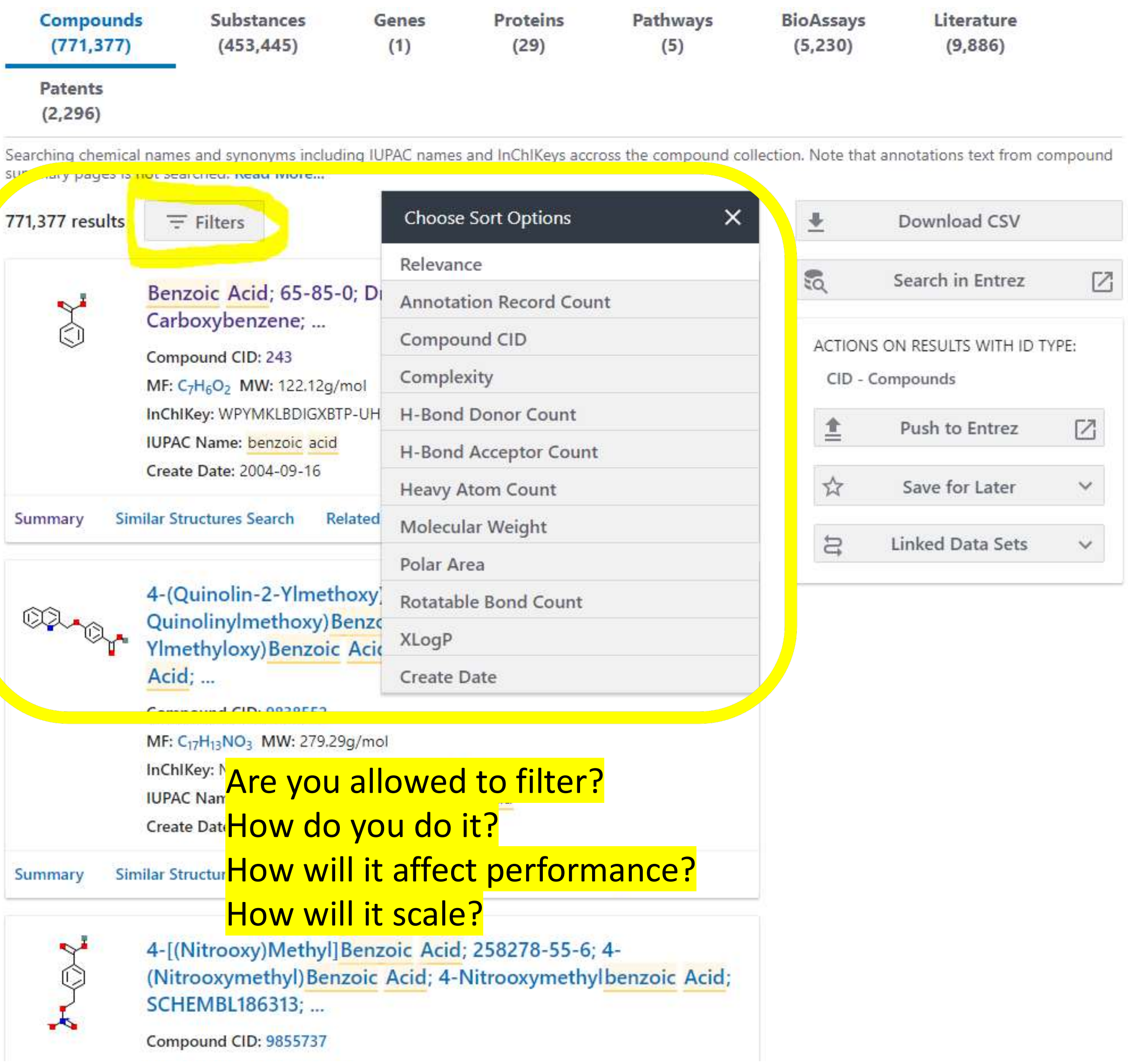
AOYSA-N

What can you afford to compute?
Every feature has a cost?
Which do you keep?
What can you do with the result?

PubChem Search

https://pubchem.ncbi.nlm.nih.gov/

- Single box, many query types
  - Chemical name, CAS#
  - Gene symbol/name
  - Molecular Formula
  - SMILES, InChI
  - SMARTS (substructure)
  - …
- Draw a structure
  - Or upload a file
- Chemical Search by
  - Identity, similarity, substructure, superstructure, mol. formula
- Upload an ID list
- Many collections
  - Compounds, Substances, BioAssays, Genes, Proteins, Pathways, Literature, Patents



Searching chemical names and synonyms including IUPAC names and InChIKeys accross the compound collection. Note that annotations text from compound su[illegible] pages is not searched. Read More...

771,377 results

Benzoic Acid; 65-85-0; Dı
Carboxybenzene; ...

Compound CID: 243
MF: $C_7H_6O_2$ MW: 122.12g/mol
InChIKey: WPYMKLBDIGXBTP-UH
IUPAC Name: benzoic acid
Create Date: 2004-09-16

4-(Quinolin-2-Ylmethoxy)
Quinolinylmethoxy)Benzo
Ylmethyloxy)Benzoic Acio
Acid; ...

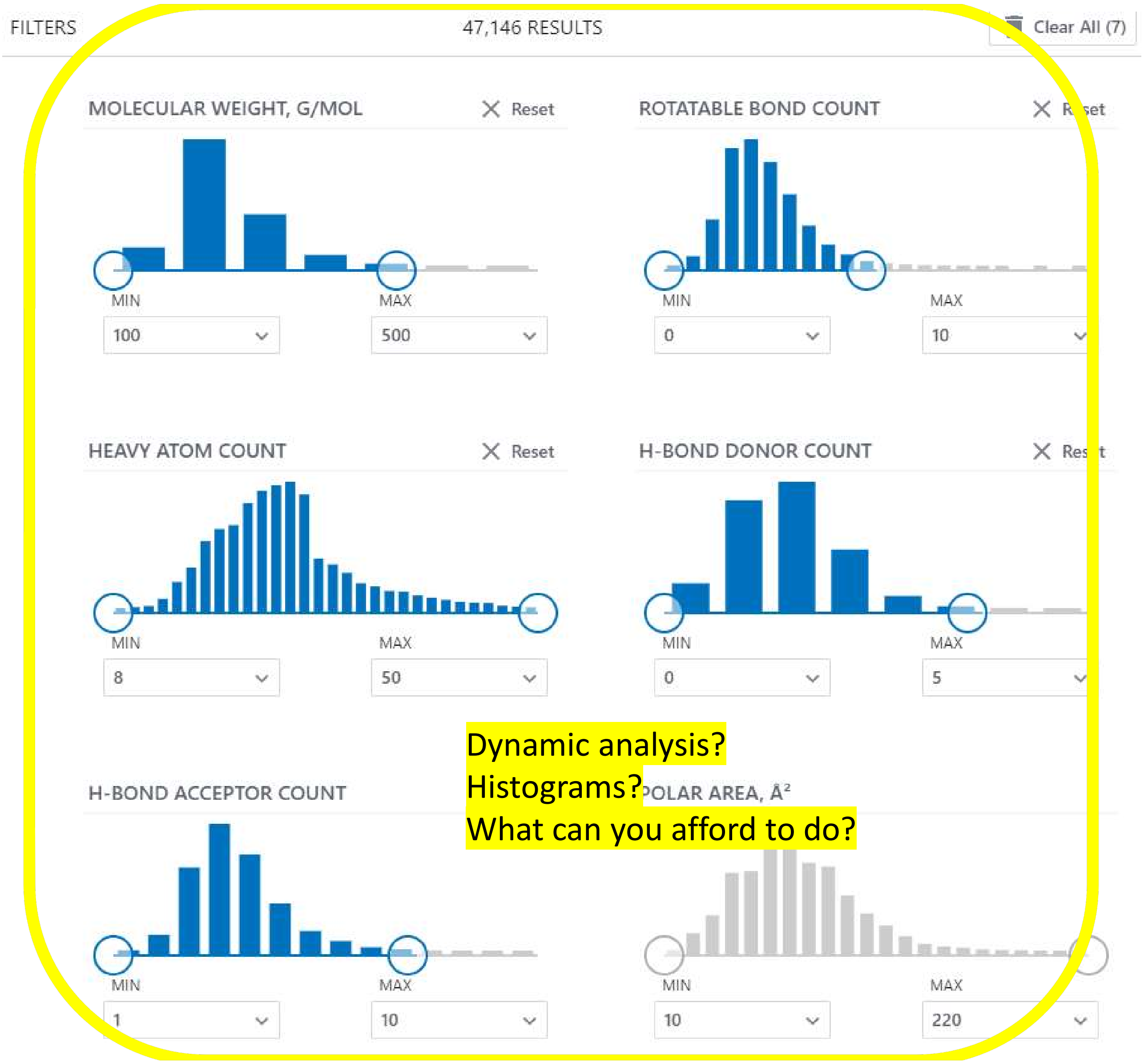
Compound CID: 9838553
MF: $C_{17}H_{13}NO_3$ MW: 279.29g/mol
InChIKey: N
IUPAC Nam
Create Dat

4-[(Nitrooxy)Methyl]Benzoic Acid; 258278-55-6; 4-(Nitrooxymethyl)Benzoic Acid; 4-Nitrooxymethylbenzoic Acid; SCHEMBL186313; ...

Compound CID: 9855737

Are you allowed to filter?
How do you do it?
How will it affect performance?
How will it scale?

- Single text box, many query types
  - Chemical name, CAS#
  - Gene symbol/name
  - Molecular Formula
  - SMILES, InChI
  - SMARTS (substructure)
  - …
- Draw a structure
  - Or upload a file
- Chemical Search by identity, similarity, substructure, superstructure, mol. formula
- Upload an ID list
- Many collections
  - Compounds, Substances, BioAssays, Genes, Proteins, Pathways, Literature, Patents



FILTERS 47,146 RESULTS Clear All (7)

MOLECULAR WEIGHT, G/MOL × Reset
MIN 100 MAX 500

ROTATABLE BOND COUNT × Reset
MIN 0 MAX 10

HEAVY ATOM COUNT × Reset
MIN 8 MAX 50

H-BOND DONOR COUNT × Reset
MIN 0 MAX 5

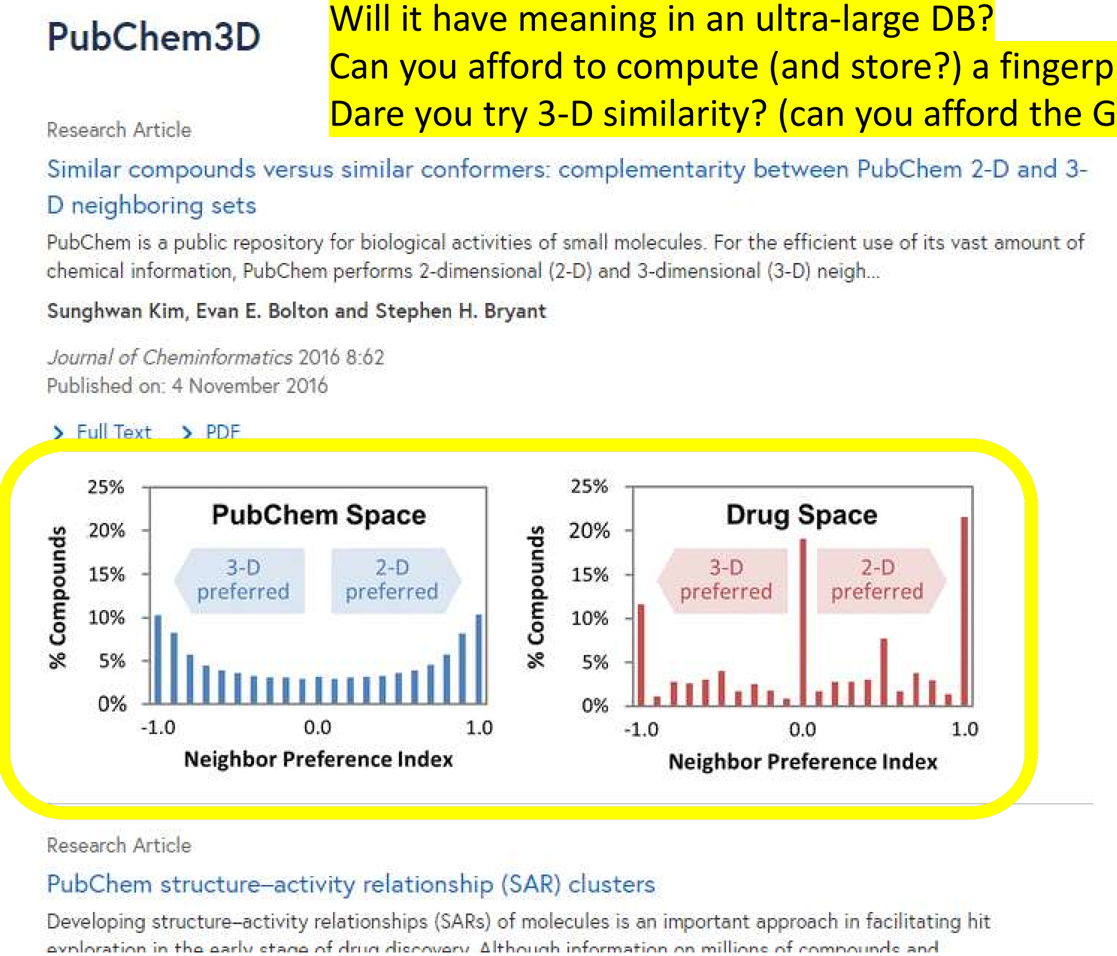
H-BOND ACCEPTOR COUNT
MIN 1 MAX 10

POLAR AREA, Å²
MIN 10 MAX 220

Dynamic analysis?
Histograms?
What can you afford to do?

- Single text box, many query types
  - Chemical name, CAS#
  - Gene symbol/name
  - Molecular Formula
  - SMILES, InChI
  - SMARTS (substructure)
  - …
- Draw a structure
  - Or upload a file
- Chemical Search by identity, similarity, substructure, superstructure, mol. formula
- Upload an ID list
- Many collections
  - Compounds, Substances, BioAssays, Genes, Proteins, Pathways, Literature, Patents

What about chemical similarity?
Will it have meaning in an ultra-large DB?
Can you afford to compute (and store?) a fingerprint?
Dare you try 3-D similarity? (can you afford the GPUs?)



PubChem3D

Research Article

Similar compounds versus similar conformers: complementarity between PubChem 2-D and 3-D neighboring sets

PubChem is a public repository for biological activities of small molecules. For the efficient use of its vast amount of chemical information, PubChem performs 2-dimensional (2-D) and 3-dimensional (3-D) neigh...

Sunghwan Kim, Evan E. Bolton and Stephen H. Bryant

Journal of Cheminformatics 2016 8:62
Published on: 4 November 2016

> Full Text > PDF

Research Article

PubChem structure–activity relationship (SAR) clusters

Developing structure–activity relationships (SARs) of molecules is an important approach in facilitating hit exploration in the early stage of drug discovery. Although information on millions of compounds and

Thematic Series of ten articles
https://www.biomedcentral.com/collections/pubchem3d

PubChem3D

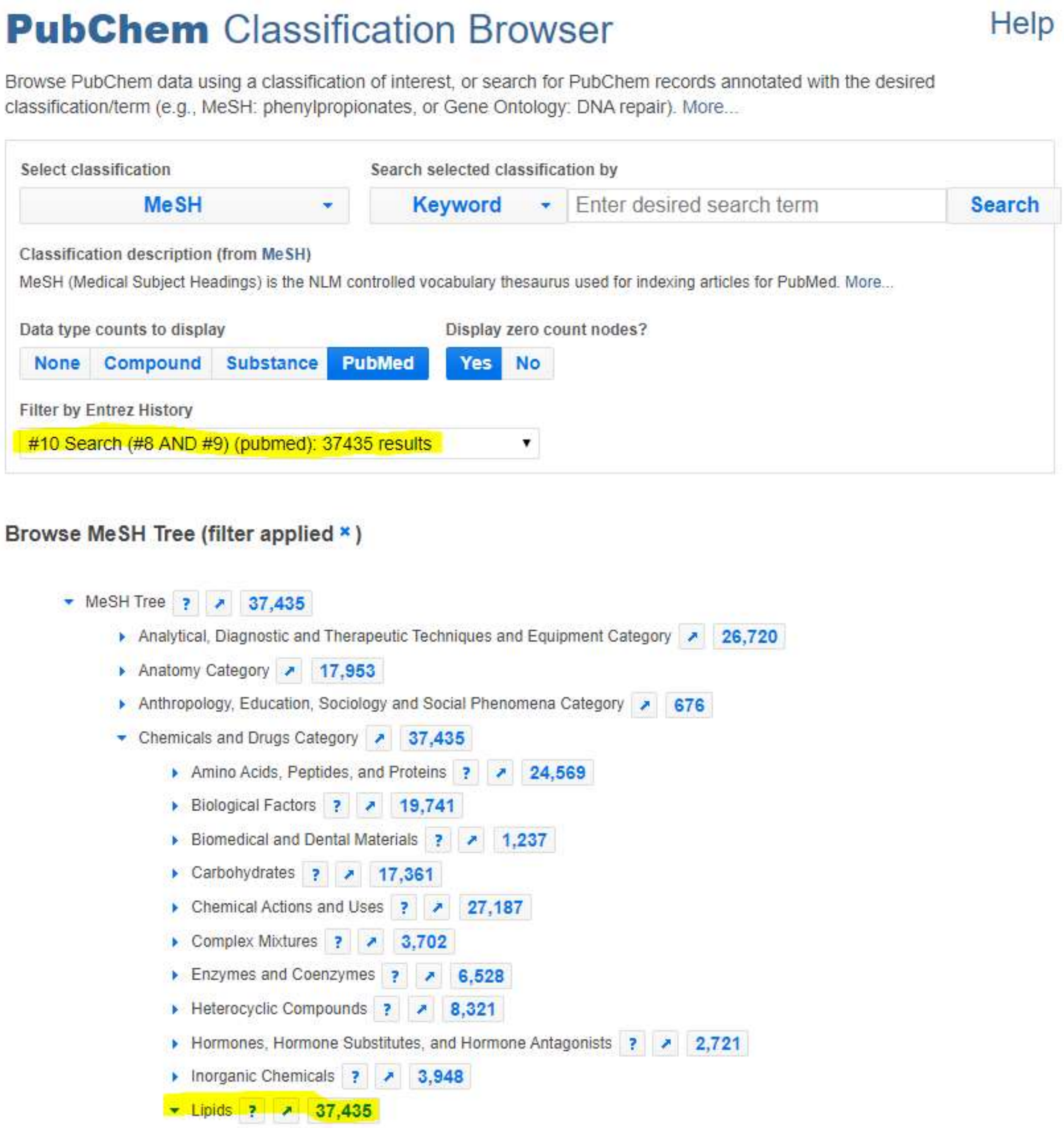
- Computationally generated 3-D structure
- Set of ten diverse conformers
- 3-D display widget
- 3-D similarity measure
  - Similar shape and protein binding features
  - Compliments 2-D similarity
- Downloadable data
  - SDF, precomputed similarity

What about analysis and subsetting?
Do you provide useful tools that help the user make sense of millions or billions of results?

Classification Browser

Browse Data

- Multiple hierarchical classifications of records in PubChem
- Includes links of different types
- Provides a way to find records with particular types of information
- Allows insightful comparisons to be made between sets of records
- Provides list import capability via Entrez (adding PubChem Search)



PubChem Classification Browser

Help

Browse PubChem data using a classification of interest, or search for PubChem records annotated with the desired classification/term (e.g., MeSH: phenylpropionates, or Gene Ontology: DNA repair). More...

Select classification: MeSH ▾

Search selected classification by: Keyword ▾ Enter desired search term Search

Classification description (from MeSH)
MeSH (Medical Subject Headings) is the NLM controlled vocabulary thesaurus used for indexing articles for PubMed. More...

Data type counts to display: None Compound Substance [✓ PubMed]

Display zero count nodes? [✓ Yes] No

Filter by Entrez History: #10 Search (#8 AND #9) (pubmed): 37435 results ▾

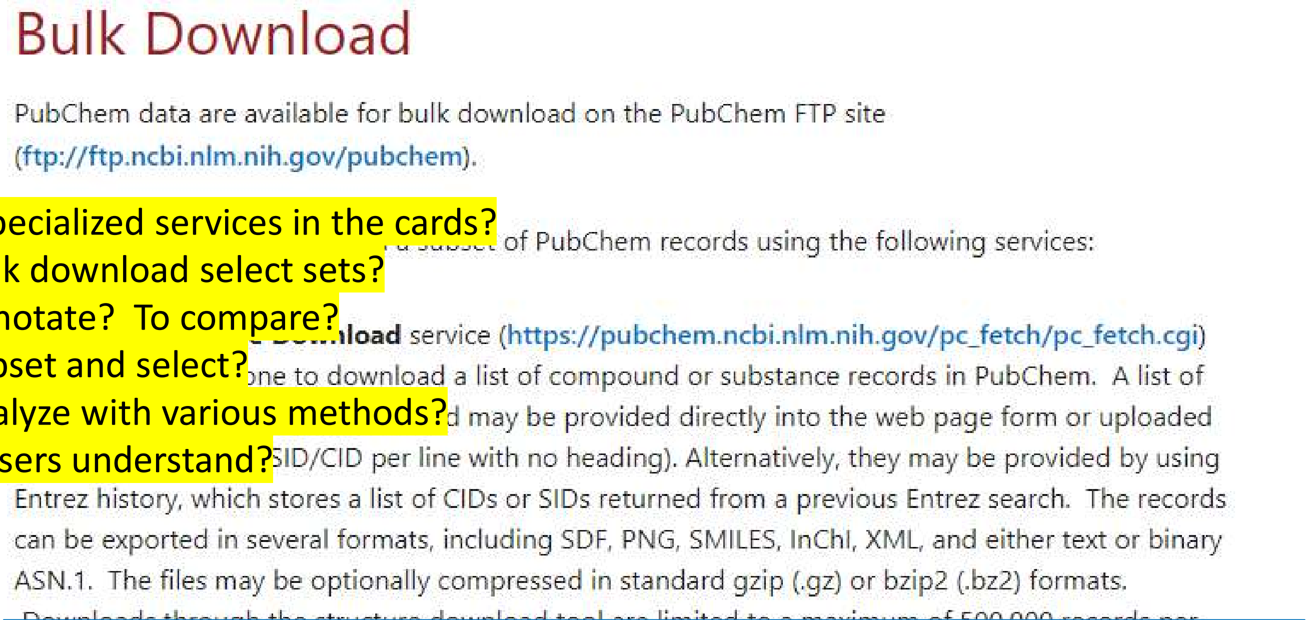
Browse MeSH Tree (filter applied ×)

- ▾ MeSH Tree ? ↗ 37,435
  - ▸ Analytical, Diagnostic and Therapeutic Techniques and Equipment Category ↗ 26,720
  - ▸ Anatomy Category ↗ 17,953
  - ▸ Anthropology, Education, Sociology and Social Phenomena Category ↗ 676
  - ▾ Chemicals and Drugs Category ↗ 37,435
    - ▸ Amino Acids, Peptides, and Proteins ? ↗ 24,569
    - ▸ Biological Factors ? ↗ 19,741
    - ▸ Biomedical and Dental Materials ? ↗ 1,237
    - ▸ Carbohydrates ? ↗ 17,361
    - ▸ Chemical Actions and Uses ? ↗ 27,187
    - ▸ Complex Mixtures ? ↗ 3,702
    - ▸ Enzymes and Coenzymes ? ↗ 6,528
    - ▸ Heterocyclic Compounds ? ↗ 8,321
    - ▸ Hormones, Hormone Substitutes, and Hormone Antagonists ? ↗ 2,721
    - ▸ Inorganic Chemicals ? ↗ 3,948
    - ▾ Lipids ? ↗ 37,435

Many Helpful Services and Functions

- Identifier Exchange Service
- Score Matrix Service
- Standardization Service
- BioActivity Dyad pages
- Entrez Indicies and Filters
- Bulk Download facilities
- AutoComplete Service
- PubChemRDF REST

Are specialized services in the cards?
To bulk download select sets?
To annotate? To compare?
To subset and select?
To analyze with various methods?
Will users understand?

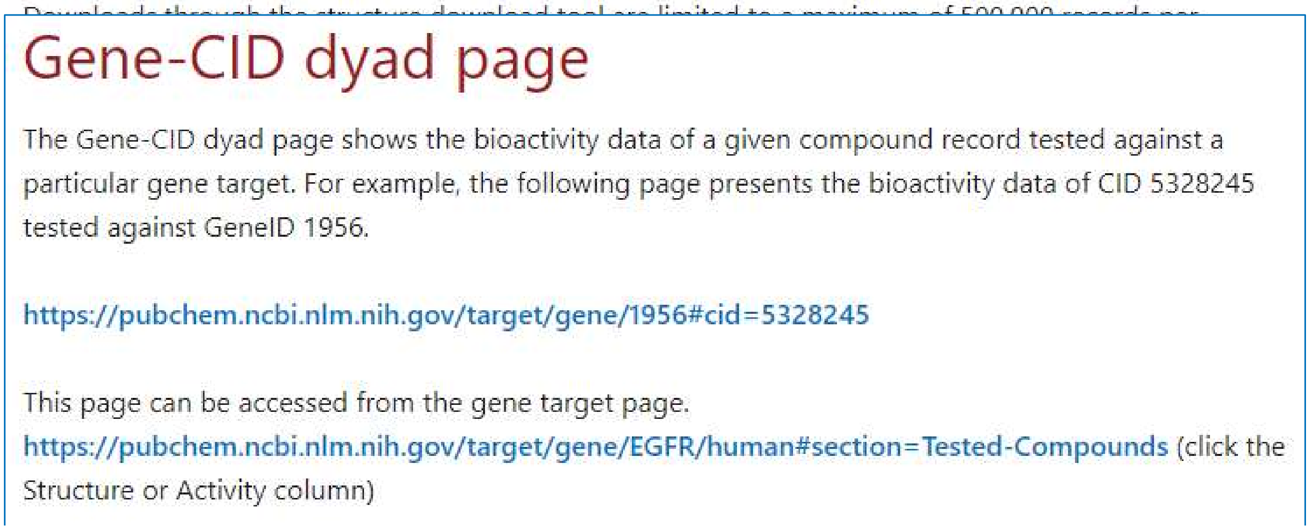


Bulk Download

PubChem data are available for bulk download on the PubChem FTP site (ftp://ftp.ncbi.nlm.nih.gov/pubchem).

of PubChem records using the following services:

load service (https://pubchem.ncbi.nlm.nih.gov/pc_fetch/pc_fetch.cgi) ne to download a list of compound or substance records in PubChem. A list of d may be provided directly into the web page form or uploaded SID/CID per line with no heading). Alternatively, they may be provided by using Entrez history, which stores a list of CIDs or SIDs returned from a previous Entrez search. The records can be exported in several formats, including SDF, PNG, SMILES, InChI, XML, and either text or binary ASN.1. The files may be optionally compressed in standard gzip (.gz) or bzip2 (.bz2) formats. Downloads through the structure download tool are limited to a maximum of 500,000 records per



Gene-CID dyad page

The Gene-CID dyad page shows the bioactivity data of a given compound record tested against a particular gene target. For example, the following page presents the bioactivity data of CID 5328245 tested against GeneID 1956.

https://pubchem.ncbi.nlm.nih.gov/target/gene/1956#cid=5328245

This page can be accessed from the gene target page. https://pubchem.ncbi.nlm.nih.gov/target/gene/EGFR/human#section=Tested-Compounds (click the Structure or Activity column)

NIH U.S. National Library of Medicine
National Center for Biotechnology Information

Lots of precedence to reduce complexity in structural representation

Biologics

1.5M chemicals are ‘biologics’ in PubChem

- Glycan, amino acid, nucleic acid monomers
- Handles substitutions and chemical linkers

NCBI Glycans page

- Symbolic Nomenclature for Glycans (SNFG)
- Working with Glycans community

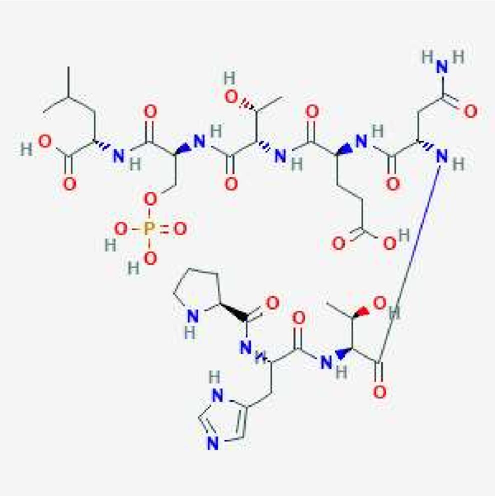
What is a glycan?

- WURCS collaboration

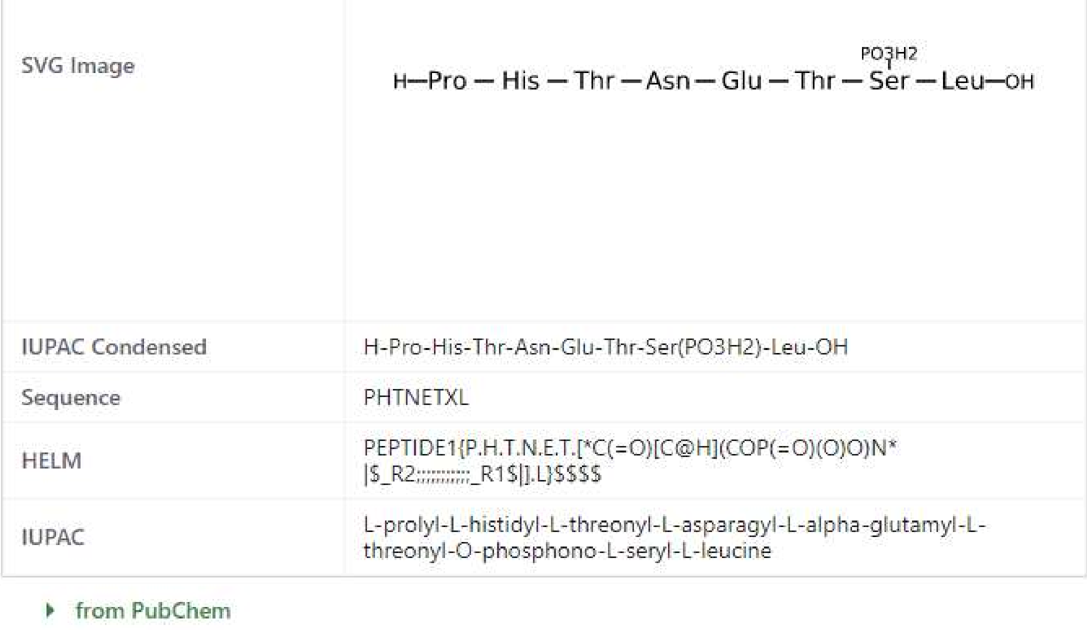
What is a biopolymer monomer?

- Pistoia Alliance HELM / EBI collaboration

U.S. National Library of Medicine
National Center for Biotechnology Information



2 Biologic Description



SVG Image	PO3H2 H—Pro — His — Thr — Asn — Glu — Thr — Ser — Leu—OH
IUPAC Condensed	H-Pro-His-Thr-Asn-Glu-Thr-Ser(PO3H2)-Leu-OH
Sequence	PHTNETXL
HELM	PEPTIDE1{P.H.T.N.E.T.[*C(=O)[C@H](COP(=O)(O)O)N* |$_R2;;;;;;;;;;;_R1$|].L}$$$$
IUPAC	L-prolyl-L-histidyl-L-threonyl-L-asparagyl-L-alpha-glutamyl-L-threonyl-O-phosphono-L-seryl-L-leucine

▸ from PubChem

NCBI

New approaches needed with ultra-large databases

- Compact formats that resist enumeration
- Computational efficiency like never before
  - Computed property computation
  - How to establish links/relationships between entities?
- Rethinking the ‘user’ workflow
  - Machines/scripts vs. humans
- Cost efficiency
  - Expensive national resource or on each user desktop?

Parting thoughts

- Very large databases (on the order of 1B-10B) are here
- Ultra-large databases (on the order of 100B-1TB) are just around the corner
- Much to do to scale
- Not every database needs or desires features found in PubChem but they will need some key features to be useful/relevant to end users
- Substantial investment in developing new algorithms, software, and rethinking databases will be necessary for practical utilization (features, cost, time)

Image credit:
https://www.myhubintranet.com/wp-content/uploads/2017/04/improving-employee-engagement.png

PubChem Crew …

Evan Bolton

Jie Chen

Tiejun Cheng

Asta Gindulyte

Jane He

Siqian He

Sunghwan Kim

Qingliang Li

Ben Shoemaker

Paul Thiessen

Bo Yu

Leonid Zaslavsky

Jian Zhang

Special thanks to the NCBI Help Desk, especially Rana Morris, and past PubChem group members

Special thanks

- Software collaborators
  - NextMove Software (Roger Sayle, John May) plus Daniel Lowe, Noel O’Boyle
  - Xemistry GmbH (Wolf D. Ihlenfeldt)
  - OpenEye Scientific Software
- PubChemRDF Collaborators
- BioHackathon (2014-2019) attendees and organizers
- All PubChem Contributors and Collaborators

- This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

[save the world •
love the people •
be happy]

Your questions?

Image credit:
https://img00.deviantart.net/7cdd/i/2007/183/8/b/science_lab_illustration_by_melodicinterval.jpg